

NOC-REK: Novel Object Captioning with Retrieved Vocabulary from External Knowledge Supplementary Material

I. Visualization of external knowledge

We show full visualization of our collected external knowledge in Fig. A. We see that our external knowledge not only adequately clusters the vocabulary (black texts), but also locates new vocabulary (red and blue texts) into appropriate clusters.

II. More examples

We show more examples of generated captions on held-out COCO dataset [4] in Figs. B, C, D, and E. We can see that our method successfully retrieves novel objects and includes them in the captions in a sensible fashion. Meanwhile, NOC [5] usually fails to generate a caption with a novel object or generates pretty weird captions. Furthermore, in Fig. C (right), the ground truth captions sometimes include *racquet*, explaining the mismatching between our method and ground truth dataset as we mentioned in main manuscript.

We further show generated captions obtained by NOC-REK and VinVL+VIVO [11, 12] on Nocaps dataset [24] in Figs. F and G. Occasionally, VinVL+VIVO [11, 12] cannot generate captions consisting of novel objects or a complete sentence. Generally, our method is capable of generating captions with more objects than those by VinVL+VIVO [11, 12].

References

- [4] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, S. Kate, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in CVPR, 2016.
- [5] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, “Captioning images with diverse objects,” in CVPR, 2017.
- [11] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, “Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training,” in AAAI, 2021.
- [12] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang,

Y. Choi, and J. Gao, “Vinvl: Making visual representations matter in vision-language models,” in CVPR, 2021.

- [24] H. Agrawal, P. Anderson, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, and S. Lee, “nocaps: novel object captioning at scale,” in ICCV, 2019.

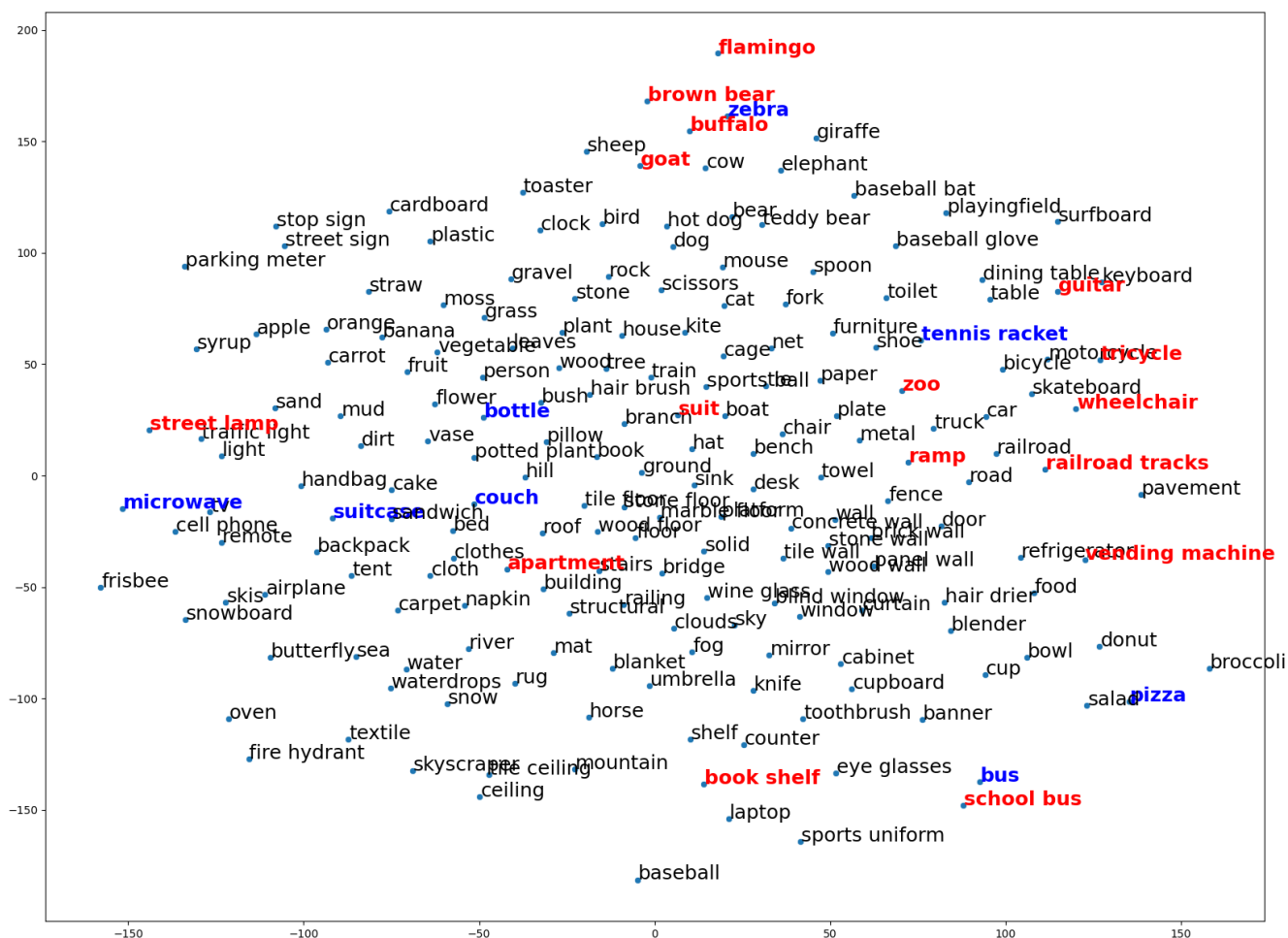


Figure A. Visualization of external knowledge using t-SNE. We see that the related vocabulary (we use objects from the seen classes of held-out COCO dataset) fall in the same cluster (black text). When we add more objects from novel classes of held-out COCO dataset (blue text) and some classes of Nocaps dataset (red text), the novel vocabulary are located at the appropriate cluster.


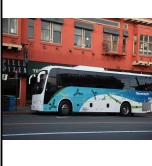






	<p>GT: A wooden table with two hotdogs in wrapping and a bottle of soda..</p> <p>NOC: A sandwich with a paper roll sitting on a table.</p> <p>NOC-REK*: A close up of a hot dog and a bottle on a table (<i>hot dog, wooden table, paper, food</i>)</p> <p>NOC-REK: A hotdog and a bottle of soda on a table. (<i>bottle, hot dog, table, paper, sandwich</i>)</p> <p>GT: Plates of food with fruit and sandwiches and bottles of beer.</p>		<p>GT: A bus stopped in front of a tall red building.</p> <p>NOC: A bus stopped at a bus stop on a city street.</p> <p>NOC-REK*: A bus parked in front of a building. (<i>bus, building, street, window, tire</i>)</p> <p>NOC-REK: a blue and white bus parked in front of a red building. (<i>city, bus, building, street, window</i>)</p> <p>GT: A red and white bus parked under an over pass.</p>
	<p>NOC: A plate of food with a hot dog and a glass of wine.</p> <p>NOC-REK*: A couple of plates of food on a table. (<i>plate, food, dining table, pole, fork</i>)</p> <p>NOC-REK: A plate of food and bottles of beer on a counter. (<i>food, bottle, beer, fork, fruit</i>)</p> <p>GT: A microwave and some bottles on a counter.</p>		<p>NOC: A bus driving past a red double decker bus.</p> <p>NOC-REK*: A public transit bus on a city street. (<i>bus, street, stop sign, car, building</i>)</p> <p>NOC-REK: A red and white bus driving under a bridge. (<i>bus, traffic light, stop sign, bridge, street</i>)</p> <p>GT: Four coach buses parked in a parking lot.</p>
	<p>NOC: A kitchen with a wooden counter top and a stove.</p> <p>NOC-REK*: A kitchen counter with a toaster oven (<i>oven, kitchen, vase, curtain, bottle</i>)</p> <p>NOC-REK: A kitchen counter with a microwave and bottles of wine. (<i>microwave, bottle, kitchen, wooden table, kitchen appliance</i>)</p> <p>GT: A man sitting at a picnic table studying a bottle.</p>		<p>NOC: A large white and red bus driving down a street.</p> <p>NOC-REK*: A row of buses parked next to each other in a parking lot. (<i>bus, street, parking lot, sky, cloud</i>)</p> <p>NOC-REK: A group of buses parked in a parking lot. (<i>bus, parking lot, building, window, mountain</i>)</p> <p>GT: A row of parked motorcycles sitting next to a white double decker bus.</p>
	<p>NOC: A man sitting on a lawn chair in front of a lawn chair.</p> <p>NOC-REK*: A man sitting at a table with a bottle of water in front of him. (<i>chair, bottle, man, grass, tree</i>)</p> <p>NOC-REK: A man sitting at a picnic table with a bottle of water. (<i>man, table, bottle, wine glass, cloth</i>)</p>		<p>NOC: A group of motorcycles parked in a row.</p> <p>NOC-REK*: A motorcycles parked in front of a bus. (<i>motorcycle, bus, person, shirt, sand</i>)</p> <p>NOC-REK: A group of motorcycles parked in front of a bus. (<i>person, motorcycle, bus, school bus, windshield</i>)</p>

Figure B. Examples of generated captions by compared methods on held-out COCO, specifically, **bottle** and **bus** classes. We show the ground-truth captions (GT) for reference. NOC [5] usually fails to generate captions with novel objects. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.









 <p>GT: A woman sitting on the couch with her laptop.</p> <p>NOC: A man is sitting on a bed with a laptop.</p> <p>NOC-REK*: A woman sitting on a couch using a laptop computer. (<i>woman, laptop, couch, book, book shelf</i>)</p> <p>NOC-REK: A woman sitting on a couch with a laptop. (<i>laptop, woman, couch, map, book shelf</i>)</p>	 <p>GT: A strainer sitting next to a microwave and a window.</p> <p>NOC: A bathroom with a toilet and a sink.</p> <p>NOC-REK*: A blue bowl sitting on top of a wooden table. (<i>wooden table, bowl, wall, window, oven</i>)</p> <p>NOC-REK: A blue bowl sitting on a table next to a microwave. (<i>bowl, microwave, table, oven, wall</i>)</p>
 <p>GT: 2 men sit on the couch, video game controllers in hands.</p> <p>NOC: A man sitting at a table eating a meal.</p> <p>NOC-REK*: Two men sitting on a couch playing a video game. (<i>people, controller, wii, couch, human face</i>)</p> <p>NOC-REK: A couple of men sitting on a couch playing a video game. (<i>man, people, couch, controller, window</i>)</p>	 <p>GT: A sofa chair and a microwave sitting on top of it.</p> <p>NOC: A large white and grey brick building with a large square bush.</p> <p>NOC-REK*: A chair sitting in front of a brick wall. (<i>chair, brick, wall, window, microwave</i>)</p> <p>NOC-REK: A green chair with a microwave on top of it. (<i>couch, chair, microwave, table, wall</i>)</p>
 <p>GT: A living room with a brown couch by a big window.</p> <p>NOC: A room with a couch, chair, table, and a tv.</p> <p>NOC-REK*: A living room filled with furniture and a large window. (<i>furniture, window, curtain, light, living room</i>)</p> <p>NOC-REK: A living room with a couch and a large window. (<i>living room, couch, window, curtain, book</i>)</p>	 <p>GT: A brown cat is sitting on top of a microwave.</p> <p>NOC: A cat is sitting on a kitchen counter.</p> <p>NOC-REK*: A cat sitting on top of a kitchen counter. (<i>cat, cabinet, wall, kitchen, microwave</i>)</p> <p>NOC-REK: A cat sitting on top of a microwave. (<i>cat, kitchen, cabinet, microwave, bottle</i>)</p>
 <p>GT: A black cat resting himself on a couch.</p> <p>NOC: A black and white cat is sitting on a black chair.</p> <p>NOC-REK*: A black cat sitting on top of a white couch. (<i>cat, couch, wall, sky, chair</i>)</p> <p>NOC-REK: A black cat laying on a couch. (<i>cat, couch, wall, floor, chair</i>)</p>	 <p>GT: Small kitchen with wood cabinets, white refrigerator, microwave.</p> <p>NOC: A kitchen with a sink and refrigerator in it.</p> <p>NOC-REK*: A white refrigerator freezer sitting inside of a kitchen. (<i>refrigerator, kitchen, cabinet, cloth, kitchen appliance</i>)</p> <p>NOC-REK: A small kitchen with a refrigerator and a microwave. (<i>kitchen, microwave, refrigerator, utensil, kitchen appliance</i>)</p>

Figure C. Examples of generated captions by compared methods on held-out COCO, specifically, **couch** and **microwave** classes. We show the ground-truth captions (GT) on for reference. NOC [5] usually fails to generate captions with novel objects. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.

 <p>GT: Two girls standing in a store eating pizza and food.</p> <p>NOC: A woman holding a cell phone standing next to a woman.</p> <p>NOC-REK*: a couple of women standing next to each other eating pizza. (<i>pizza, woman, people, sign, refrigerator</i>)</p> <p>NOC-REK: a couple of women eating pizza in a store. (<i>pizza, woman, people, store, sign</i>)</p>	 <p>GT: A man holding a tennis racket and playing tennis on a tennis court.</p> <p>NOC: A man is standing in the grass with a racket.</p> <p>NOC-REK*: A man holding a tennis racket on top of a tennis court. (<i>sky, person, racket, tree, fence</i>)</p> <p>NOC-REK: A man holding a tennis racket in front of a fence. (<i>racket, man, sunglasses, fence, shirt</i>)</p>
 <p>GT: Two people sitting behind a large pizza on a table.</p> <p>NOC: Pizza sitting on a plate on a table.</p> <p>NOC-REK*: A couple of people sitting at a table with a pizza. (<i>people, pizza, couch, wall, window</i>)</p> <p>NOC-REK: A couple of people sitting in front of a large pizza. (<i>pizza, people, person, table, couch</i>)</p>	 <p>GT: A woman standing on a tennis court holding a racquet.</p> <p>NOC: A tennis player is getting ready to return a ball.</p> <p>NOC-REK*: A woman in a short skirt holding a tennis racket. (<i>racket, woman, skirt, pole, sand</i>)</p> <p>NOC-REK: A woman in a white dress holding a tennis racket. (<i>racket, tennis court, woman, dress, skirt</i>)</p>
 <p>GT: Someone cutting a small pizza with a pizza cutter.</p> <p>NOC: A person cutting a piece of cake on a cutting board.</p> <p>NOC-REK*: A person cutting a pizza. (<i>pizza, person, table, knife, plate</i>)</p> <p>NOC-REK: A person cutting a pizza with a pizza cutter. (<i>pizza, cutter, person, table, plate</i>)</p>	 <p>GT: A group of women sitting around each other holding racquets.</p> <p>NOC: A woman sitting on a chair with a teddy bear.</p> <p>NOC-REK*: A group of women sitting next to each other on a bench. (<i>woman, person, people, racket, bench</i>)</p> <p>NOC-REK: A group of women sitting on a bench with tennis rackets. (<i>racket, people, bench, woman, wall</i>)</p>
 <p>GT: A young girl sitting at a table that has two pizzas and a peps beverage glass on it.</p> <p>NOC: A woman sitting on a table next to a plate of food.</p> <p>NOC-REK*: A little girl sitting at a table with a plate of pizza. (<i>girl, chair, pizza, food, window</i>)</p> <p>NOC-REK: A little girl sitting at a table with two pizzas. (<i>girl, pizza, plate, glass, table</i>)</p>	 <p>GT: A young girl lays on a wooden floor while clutching a tennis racquet.</p> <p>NOC: A woman is holding a racket in her hand.</p> <p>NOC-REK*: A young girl laying on a tennis court holding a tennis racket. (<i>woman, shirt, racket, girl, tennis court</i>)</p> <p>NOC-REK: A woman laying on the floor with a tennis racket. (<i>woman, floor, racket, hat, coat</i>)</p>

Figure D. Examples of generated captions by compared methods on held-out COCO, specifically, **pizza** and **racket** classes. We show the ground-truth captions (GT) for reference. NOC [5] usually fails to generate captions with novel objects. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.

 <p>GT: A cat is sitting on top of a suitcase.</p> <p>NOC: A cat laying on top of a suitcase.</p> <p>NOC-REK*: A cat sitting on top of a blue piece of luggage. (<i>kitty, cat, luggage, floor, wheel</i>)</p> <p>NOC-REK: A cat sitting on top of a blue suitcase. (<i>cat, suitcase, luggage, floor, wooden floor</i>)</p>	 <p>GT: A group of zebra's eating hay from a trough.</p> <p>NOC: Zebra grazing in a fenced area next to a zebra.</p> <p>NOC-REK*: A herd of zebra standing on top of a dry grass field. (<i>zebra, field, tree, grass, hill</i>)</p> <p>NOC-REK: A group of zebras eating hay in a field. (<i>zebra, hay, grass, field, tree</i>)</p>
 <p>GT: A man sitting on a wooden bench with a suitcase and a box on his lap.</p> <p>NOC: A man sitting on a bench with a dog.</p> <p>NOC-REK*: A man sitting on a bench next to a suitcase. (<i>bench, man, person, suitcase, people</i>)</p> <p>NOC-REK: A man sitting on a bench with a suitcase. (<i>suitcase, man, bench, suit, brick</i>)</p>	 <p>GT: A young man standing in front of a zebra in an enclosure.</p> <p>NOC: A man standing next to a zebra standing next to a wooden fence.</p> <p>NOC-REK*: A man standing next to a zebra on a field. (<i>man, zebra, fence, tree, field</i>)</p> <p>NOC-REK: A man standing in front of a zebra at a zoo. (<i>zebra, man, zoo, grass, tree</i>)</p>
 <p>GT: A suitcase lying on the ground with a spare tire and jack.</p> <p>NOC: A black and white cat is sitting on a car.</p> <p>NOC-REK*: A couple of bags of luggage sitting on top of a gravel ground. (<i>luggage, wheel, tire, ground, pole</i>)</p> <p>NOC-REK: A black suitcase sitting next to a tire. (<i>suitcase, luggage, tire, wheel, stone</i>)</p>	 <p>GT: A realistic painting of two zebras huddled in brown grass.</p> <p>NOC: Zebra standing next to another zebra standing next to another zebra.</p> <p>NOC-REK*: Two zebra standing next to each other on a dry grass field. (<i>zebra, grass, field, cloud, tree</i>)</p> <p>NOC-REK: A painting of two zebras standing next to each other. (<i>zebra, grass, painting, sky, cloud</i>)</p>
 <p>GT: A boy in a tie holding a small suitcase.</p> <p>NOC: A man standing next to a man on a skateboard.</p> <p>NOC-REK*: A little boy that is standing next to a suitcase. (<i>person, suitcase, window, floor, people</i>)</p> <p>NOC-REK: A young boy wearing a tie holding a suitcase. (<i>man, tie, suitcase, shirt, table</i>)</p>	 <p>GT: A field full of zebras and giraffes at a zoo.</p> <p>NOC: Zebras grazing near trees in a field near trees.</p> <p>NOC-REK*: A herd of zebra grazing on a lush green field. (<i>grass, field, tree, zebra, sky</i>)</p> <p>NOC-REK: A couple of zebras and giraffes grazing in a field. (<i>zebra, giraffe, grass, tree, bird</i>)</p>

Figure E. Examples of generated captions by compared methods on held-out COCO, specifically, **suitcase** and **zebra** classes. We show the ground-truth captions (GT) for reference. NOC [5] usually fails to generate captions with novel objects. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.



VinVL + VIVO: a wall with a bunch of wooden jars with **cellos** on it in a wall.

NOC-REK*: a group of **cellos** sitting on top of a table. (cello, table, violin, wall, musical instrument)

NOC-REK: a group of wooden **cello** sticks on a wall. (cello, wall, musical instrument, violin)



VinVL + VIVO: a desk with two **computer monitors** and a **microphone** next to a glass of wine.

NOC-REK*: a desk with two **computer monitors** and a **microphone**. (computer monitor, microphone, desk, lamp, bottle)

NOC-REK: a desk with a **laptop** and a **microphone** and a **computer monitor** and a **table lamp**. (microphone, computer monitor, desk, laptop, table lamp)



VinVL + VIVO: a **leopard** cat standing next to a **fence** with a **jaguar** in front of a **forest**.

NOC-REK*: a **jaguar** and a **leopard** print cat standing next to a **fence**. (leopard, jaguar, fence, cat, tree)

NOC-REK: two **jaguar leopards** standing on some **rocks** near a **fence**. (leopard, jaguar, rock, fence, forest)



VinVL + VIVO: a **camera** sitting next to a camera with a camera.

NOC-REK*: a **camera** sitting on top of a camera **tripod**. (tripod, camera, wood, door, wall)

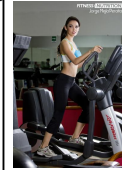
NOC-REK: a **camera** is sitting on top of a **tripod**. (tripod, camera, wall, curtain, wood)



VinVL + VIVO: a person standing in front of a blue **van** parked in a parking lot.

NOC-REK*: a blue **van** parked in a parking lot with a man. (van, man, person, wheel, parking lot)

NOC-REK: a person standing next to a blue and white **van** in a parking lot. (man, parking lot, van, tire, windshield)



VinVL + VIVO: a **woman** sitting on top of an airport stationary bicycle.

NOC-REK*: a **woman** sitting on a **treadmill** in a stationary bicycle. (person, woman, treadmill, bicycle, wall)

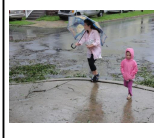
NOC-REK: a **woman** is standing on a **treadmill** in a gym. (woman, treadmill, mirror, wall, sport equipment)



VinVL + VIVO: a **flowerpot** with a houseplant with a tree in front of it.

NOC-REK*: a houseplant with a tree and a **flowerpot** on it. (tree, flowerpot, bowl, paper, wall)

NOC-REK: a small tree is houseplant in a **flowerpot**. (flowerpot, plotted plan, tree, soil)



VinVL + VIVO: A couple of children walking in the **rain** with umbrellas in a street.

NOC-REK*: A **woman** and a little **girl** walking with an umbrella. (umbrella, people, girl, sidewalk, street)

NOC-REK: A **woman** and a little **girl** walking in the **rain** with an umbrella. (umbrella, woman, girl, street, rain)

Figure F. Examples of generated captions by compared methods Nocaps. VinVL+VIVO [11, 12] sometimes cannot include the novel objects in the captions. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.

 <p>VinVL + VIVO: a couple of people standing next to each other in a field with a sign.</p> <p>NOC-REK*: a woman standing in front of a wall with a sign. (<i>sign, woman, person, wall, clothes</i>)</p> <p>NOC-REK: a person in a military uniform with a quote on the back. (<i>person, jacket, uniform, wall, sign</i>)</p>	 <p>VinVL + VIVO: a cake with sprinkles on a table with a table.</p> <p>NOC-REK*: a birthday cake with a bunch of toys on it. (<i>toy, table, birthday cake, building, human</i>)</p> <p>NOC-REK: a birthday cake covered in colorful candy on a table. (<i>birthday cake, candy, house, table, wall</i>)</p>
 <p>VinVL + VIVO: a group of wooden cabinets with a table with a table in the background.</p> <p>NOC-REK*: a couple of wooden cabinets with a window. (<i>cabinet, window, wood, table, door</i>)</p> <p>NOC-REK: two wooden cabinet doors on the side of each other. (<i>cabinet, lock, window, door, wood</i>)</p>	 <p>VinVL + VIVO: a man sitting at a desk with a computer on top of a office building.</p> <p>NOC-REK*: a man sitting at a desk with a office building. (<i>man, computer, office building, office chair, desk</i>)</p> <p>NOC-REK: a man sitting at a desk in a office building with a computer. (<i>man, office building, office supplies, computer, desk</i>)</p>
 <p>VinVL + VIVO: a smoke stack of water with a lighthouse with a fire hydrant in the background.</p> <p>NOC-REK*: a white lighthouse sitting next to a body of water. (<i>lighthouse, water, cloud, hole, pole</i>)</p> <p>NOC-REK: a lighthouse with a smoke stack in the sky. (<i>sky, lighthouse, smoke, cloud, window</i>)</p>	 <p>VinVL + VIVO: a cat laying on top of a carnivore with trees in the background.</p> <p>NOC-REK*: a carnivore cat laying on top of a tree. (<i>tree, cat, carnivore, background, dog</i>)</p> <p>NOC-REK: a brown dog laying on a carnivore in a tree. (<i>dog, carnivore, tree, cat, leopard</i>)</p>
 <p>VinVL + VIVO: a coffee cup sitting on a tea bag.</p> <p>NOC-REK*: a coffee cup sitting on top of a coffee table. (<i>coffee cup, coffee table, table, cup, window sill</i>)</p> <p>NOC-REK: a coffee cup sitting on top of a coffee table on a window sill. (<i>coffee cup, window sill, coffee table, tea, cup</i>)</p>	 <p>VinVL + VIVO: a couple of babies laying next to each other on a beach. the ocean.</p> <p>NOC-REK*: a man and a woman holding a baby. (<i>man, baby, woman, beach, sky</i>)</p> <p>NOC-REK: a man and a woman holding a baby on the beach. (<i>man, woman, sky, baby, beach</i>)</p>

Figure G. Examples of generated captions by compared methods Nocaps. VinVL+VIVO [11, 12] sometimes cannot include the novel objects in the captions. Our NOC-REK, on the other hand, successfully generates correct, fluent, and coherent captions with novel objects. Words in parentheses are top-5 retrieved vocabulary by our method that are reasonably related to objects in image. **Red** texts indicate novel objects in the captions.